Addressing the Elephant in the Room: Robust Animal Re-Identification with Unsupervised Part-Based Feature Alignment

Yingxue Yu   Vidit Vidit   Andrey Davydov   Martin Engilberge   Pascal Fua

EPFL

{firstname.lastname}@epfl.ch

Abstract

Animal Re-ID is crucial for wildlife conservation, yet it faces unique challenges compared to person Re-ID. First, the scarcity and lack of diversity in datasets lead to background-biased models. Second, animal Re-ID depends on subtle, species-specific cues, further complicated by variations in pose, background, and lighting. This study addresses background biases by proposing a method to systematically remove backgrounds in both training and evaluation phases. And unlike prior works that depend on pose annotations, our approach utilizes an unsupervised technique for feature alignment across body parts and pose variations, enhancing practicality. Our method achieves superior results on three key animal Re-ID datasets: ATRW, YakReID-103, and ELPephants.

1. Introduction

While Person Re-identification (Re-ID) has seen considerable advancements [1, 4, 6, 7, 11, 16, 20, 23, 25, 28–31, 33], animal Re-ID [13, 15, 32] remains underexplored despite its significance in both industrial and conservation contexts. Traditional animal identification methods are labor-intensive and impractical for large populations, highlighting the need for improved automated Re-ID techniques [22].

Animal Re-ID faces unique challenges, including small non-diverse datasets leading to background bias, greater variations within an individual and smaller variations between different individuals compared to humans. These challenges necessitate animal-specific Re-ID methods.

This study focuses on Re-ID for elephants, yaks, and tigers, emphasizing the importance of feature alignment across pose variations due to their four-legged nature and the variability in their appearances. Unlike prior works [14, 18, 19, 25] that depend on pose annotations, our approach employs an unsupervised learning method [27] to identify semantically similar animal parts, enhancing Re-ID performance.
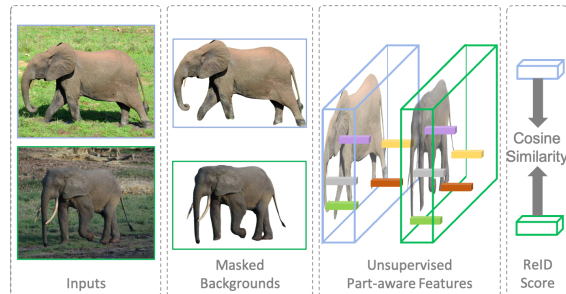

Figure 1. **Proposed Animal Re-ID Approach:** Addressing background bias in Re-ID models, our method masks out backgrounds to focus on the animal. It learns part-aware representations, ensuring consistency across subjects. Part-aware features are merged and a final Re-ID score is computed via cosine similarity.

Our contributions illustrated in Fig. 1 include: (i) Reducing background bias by creating and sharing background-free images, (ii) Utilizing an unsupervised method for learning part-based descriptors, thereby improving Re-ID accuracy without needing pose annotations and (iii) assessing the transferability of our method between different species, demonstrating the feasibility of transferring pose alignment across various species.

2. Method

We present our approach for Animal Re-ID, starting with our technique for background removal to focus evaluation on animals rather than their surroundings. Next, we outline our Re-ID model's architecture and training approach. Lastly, we detail our integration of unsupervised landmark detection with our model to enable the learning of part-aware representations.

2.1. Background and Re-ID

Background overfitting is a significant issue in animal Re-ID, exacerbated by small dataset sizes and repetitive back-

---

The code is made available in the following link: https://github.com/Chloe-Yu/Animal-Re-ID

Figure 2. **Bias Towards Background:** On the left, we display samples from the YakReID-103 dataset. Utilizing features from the PGCFL [19] model, we identify the nearest neighbors. Each image's entity label is presented in the bottom-left corner. The retrieved images showcase four distinct entities, all sharing a remarkably similar background. This indicates that distances in the feature space are significantly influenced by background similarities. On the right, we exhibit the outcomes of our proposed background segmentation protocol. The top-left image is the original, the bottom-left depicts results from SAM, the bottom-right from ISNet, and the top-right combines outputs from both SAM and ISNet.

grounds. It is particularly problematic in real-world scenarios where the same entity can be seen with lots of different backgrounds. We illustrate this problem on the left side of Fig. 2.

To counter this, we diverge from previous strategies that adapt models to ignore backgrounds. Instead, we propose altering the data by removing backgrounds from all images using segmentation models ISNet [2] and SAM [12]. We combine both models to produce refined segmentation masks, balancing ISNet's completeness with SAM's precision, particularly for distinguishing animal parts like tusks and horns. To get the best of both worlds, we combine both approaches to obtain foreground mask $M_{\mathbf{x}}$ from a given image $\mathbf{x}$ as :

$$M_{\mathbf{x}} = \bigcup \{\mathbf{m} | \mathbf{m} \in S_{SAM}(\mathbf{x}) \land h(\mathbf{m}; S_{ISNet}(\mathbf{x})) < \Theta\} ,$$
(1)

where $S_{SAM}(x)$ are processed object masks from SAM, and $S_{ISNet}(x)$ is the output mask from ISNet. h is a criterion (such as IoU) that filters $S_{SAM}(x)$ given $S_{ISNet}(x)$ with threshold Θ. We provide additional details about the criterion in the supplementary material.

The combination of both models yields better segmentation masks as shown in Fig. 2. In the following sections, we use this segmentation protocol to extract foreground masks for the different benchmarks and mask out the background both during training and evaluation. Code and processed images will be made available to facilitate replication and further research.

2.2. Re-ID Approach

Formalism

In a standard Re-ID scenario, given a query image $\mathbf{q}$ of dimensions $w \times h \times 3$, the goal is to rank a gallery set of N images $\mathcal{G} = \{\mathbf{g}_j\}_1^N$ to find those matching the query's identity. Images are ranked based on the cosine similarity between the query representation $f(\mathbf{q}) \in \mathbb{R}^d$ and each gallery image representation $f(\mathbf{g}_j) \in \mathbb{R}^d$, calculated as $c(f(\mathbf{q}), f(\mathbf{g}_j)) = \frac{\langle f(\mathbf{q}), f(\mathbf{g}_j) \rangle}{\|f(\mathbf{q})\| \|f(\mathbf{g}_j)\|}$. Here, f is the embedding function into a d-dimensional Re-ID space, parametrized by θ. The training process focuses on optimizing θ. The gallery encompasses one or multiple images of each unique animal identity, with a total of C_{id} identities ($C_{id} <= N$).

Architecture

Consistent with prior Animal Re-ID studies [19, 32], our model employs a modified SE-ResNet50 [9], following adaptations in [21]. We utilize the initial five convolutional blocks of SE-ResNet50. The output from the fifth block is processed through a Global Average Pooling (GAP) layer [17], condensing spatial dimensions into 1D vectors. These vectors are then passed through a sequence comprising a linear layer, batch normalization [10], and dropout [24], producing the Re-ID vector representation. For identity and orientation classification, the Re-ID features are further passed through two linear layers to produce the final logits.

Formally, for an input image $\mathbf{x} \in (0, 255)^{W \times H \times 3}$, our Re-ID model operates as:

$$\mathbf{x} \overset{\text{BR}}{\longmapsto} \bar{\mathbf{x}} \overset{\text{SE}}{\longmapsto} \mathbf{F} \overset{\text{GAP}}{\longmapsto} \mathbf{G} \overset{\text{FC}_{\theta_1}}{\longmapsto} f(\mathbf{x}) \overset{\text{FC}_{\theta_{2:3}}}{\longmapsto} \genfrac{}{}{0pt}{}{\hat{y}_{lr}}{\hat{y}_{id}} , \quad (2)$$

where BR denotes the background removal step, producing the background-free image $\bar{\mathbf{x}}$. SE corresponds to the first 5 layers of a pre-trained SE-ResNet50. F in $\mathbb{R}^{w \times h \times d_{se}}$ correspond to the activation of the fifth layer. G in $\mathbb{R}^{d_g}$ is the output of the Global Average Pooling operation. $f(\mathbf{x}) \in \mathbb{R}^d$ is the Re-ID representation, and $\hat{y}_{lr} \in \mathbb{R}^1$ and $\hat{y}_{id} \in \mathbb{R}^{C_{id}}$
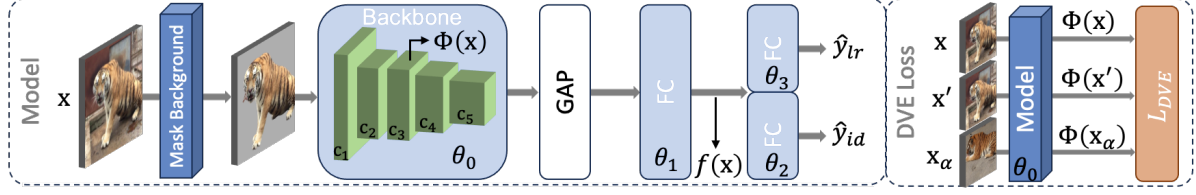
Figure 3. **Overall architecture:** Left - proposed model's architecture. Input is a single image, processed through the first 3 layers of backbone and a convolutional block to extract DVE features $\Phi(\mathbf{x})$. Features produced after the fifth backbone layer continue to global average pooling and a linear layer for Re-ID features $f(\mathbf{x})$, then pass through two classification heads for ID class and orientation prediction via linear layers and softmax operations. Right - L_{DVE} for part-aware representation.

are orientation and identity classification logits respectively. The complete architecture is shown in Fig. 3.

2.3. Part-aware Feature Learning

Identifying individuals within the same species depends on fine body part details, such as tigers' stripes [8], yaks' horns and fur, or elephants' ears and tusks. Part-based Re-ID methods, which are effective [15, 18], usually need pose annotations. Our unsupervised method leverages the Descriptor Vector Exchange (DVE) for learning part-specific features without posture labels, using DVE's technique for unsupervised dense landmark prediction [27].

DVE's local image descriptors are designed to be equivariant to transformations and invariant to variations within a category. The DVE objective function is:

$$L_{DVE} = \frac{1}{|\Omega|^2} \int_\Omega \int_\Omega \|v - gu\| p(v|u; \Phi, \mathbf{x}, \mathbf{x}', \mathbf{x}_\alpha) du dv \ , \tag{3}$$

where Φ is a projection from the image domain to the local descriptors domain, g is a random warping function, $\mathbf{x}$ is an image, and $\mathbf{x}' = g\mathbf{x}$ its deformation. $p(v|u; \Phi, \mathbf{x}, \mathbf{x}', \mathbf{x}_\alpha)$ is the probability of pixel u in image $\mathbf{x}$ matching with pixel v in image $\mathbf{x}'$. The probability computation uses an auxiliary image $\mathbf{x}_\alpha$ to make the local descriptor invariant to intra-category variations. For more information about the DVE objective function please refer to [27].

DVE for Re-ID DVE descriptors are invariant to intra-category variations, meaning similar parts across different subjects (e.g. left leg of a tiger) will have comparable descriptors. Leveraging this, we incorporate the DVE property into our Re-ID model by using the DVE loss L_{DVE} to refine our model's features. Specifically, the activation of the third convolutional layer of the SE-ResNet is fed into an extra convolution layer to get the descriptors $\Phi(\mathbf{x})$ on which L_{DVE} is applied.

Formally, those local descriptors are the results of the following steps:

$$\mathbf{x} \xmapsto{BR} \bar{\mathbf{x}} \xmapsto{SE_{c_{1:3}}} \mathbf{F}' \xmapsto{CONV} \Phi(\bar{\mathbf{x}}) \ , \tag{4}$$

$\mathbf{F}'$ in $\mathbb{R}^{w' \times h' \times d'_{se}}$ correspond to the activation of the third layer. $\Phi(\bar{\mathbf{x}})$ in $\mathbb{R}^{w' \times h' \times d_{dve}}$ is the DVE feature.

The output resolution of our model's backbone is reduced by a factor of 8 compared to the input image. Given that the resolution post-5th layer is too low for effective local descriptor learning via DVE, we opt to apply the DVE objective at a higher resolution stage, specifically after the third layer of the SE-ResNet, where the output is only downscaled by a factor of 4. Fig. 3 illustrates how L_{DVE} is obtained.

2.4. Loss Function and Training

In addition to L_{DVE}, our model incorporates standard Re-ID losses similar to those in prior studies [19, 32]. The loss function comprises two classification losses—one for orientation and another for identity—along with a loss for learning the Re-ID representation. The classification losses read as follows:

$$L_{ID} = -\sum_{c=1}^{C_{id}} y_{id}^c \log(\hat{y}_{id}^c) \ , \tag{5}$$

$$L_{LR} = -(y_{lr} \log(\hat{y}_{lr}) + (1 - y_{lr}) \log(1 - \hat{y}_{lr})) \ , \tag{6}$$

where C_{id} is the number of entity classes, y_{id} and y_{lr} is the ground truth label for entity and orientation, respectively.

For learning the Re-ID representation, we diverge from the common use of triplet loss in previous studies and instead utilize circle loss [26]. Given a sample $\mathbf{x}$, let's consider K within-class and L between-class similarity scores, denoted by $s_p^i (i = 1, 2, ..., K)$ and $s_j^n (j = 1, 2, ..., L)$, respectively. The formulation of the circle loss is as follows:

$$L_{reID} = \log \left[\sum_{j=1}^{L} \exp \left(\gamma [s_n^j + m]_+ \times (s_n^j - m) \right) \right.$$
$$\left. \times \sum_{i=1}^{K} \exp \left(-\gamma [-s_p^i + 1 + m]_+ \times (s_p^i - 1 + m) \right) \right] , \tag{7}$$

3

| Masked Backgrounds | | ATRW | | |
Train	Test	mmAP	R@1(s)	R@1(c)
		74.5	95.7	90.3
	✓	60.2	88.6	83.4
✓	✓	66.9	90.8	86.3
✓		73.4	96.9	89.7

Table 1. **Background Bias Measurement** Using the ATRW dataset, we assessed background bias in the PGCFL re-ID model. Training on original images led to performance collapse with masked backgrounds, highlighting benchmark bias and its effect on the model. Training on images with masked backgrounds improved model performance, which further enhanced with background inclusion.

where γ is the scale factor and m is the margin.

The overall objective function used during training is as follows:

$$L = L_{ID} + L_{LR} + \lambda_{reID} \times L_{reID} + \lambda_{DVE} \times L_{DVE} \ , \ (8)$$

where λ_{reID} and λ_{DVE} are hyperparameters allowing to tune respectively the strength of L_{reID} and of the regularization effect of L_{DVE}.

3. Experiments

In the following, we first describe the datasets, metrics and training procedures. Then, we investigate the effect of background removal on Re-ID performance. Finally, we show the benefit of the proposed part-aware features through a series of quantitative and qualitative evaluations.

Datasets The ATRW dataset [15] is the largest wildlife Re-ID dataset, featuring 182 entities across 92 tigers, with a training set of 1,887 images from 107 entities and a test set of 1,762 images from 75 entities, without utilizing provided pose annotations. YakReID-103 [32] includes 1,404 training images of 121 entities, and we only use the hard-testing subset of 433 images, where similar poses and backgrounds are excluded. ELPephants [13] contains 2,078 images of 276 elephants. We manually completed the dataset's partial heading direction annotations and included only entities with multiple side-view images. It's split into 1,380 training and 380 testing images without predefined bounds. Each dataset ensures no entity overlap between training and testing, considering each side of an individual as a distinct entity.

Metrics The evaluation employs mean average precision (mAP) and recall at 1 (R@1), with AP calculated per query
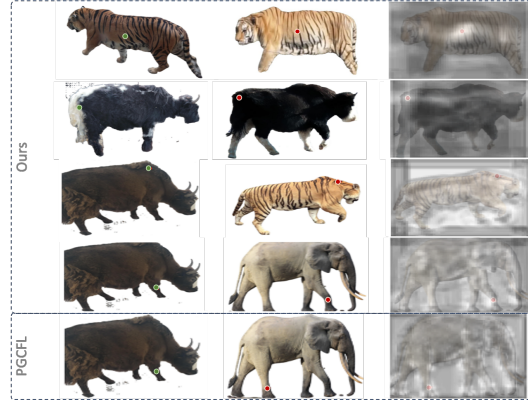

Figure 4. **Visualization of the feature learned with** L_{DVE} The first two rows show intra-species part alignment, the next two rows demonstrate that a model trained *solely* on tiger can generalize to other species and maintain alignment even in inter-species scenario. The final row is the results from the PGCFL baseline. In each row, the green dot in the left image is the local query, while the red dot in the center image indicates its matching point. The right-most image provides a heatmap overlaid on the target image, showcasing the similarities between the local query and the center image.

from ranked gallery lists. For ATRW, metrics are separately computed for single-camera (R@1(s)) and cross-camera (R@1(c)) settings, with mmAP as their average. YakReID-103 and ELPephants lack camera data, considering any same-entity gallery image as positive.

Implementation details Utilizing SE-ResNet50 pretrained on ImageNet as the backbone, similar to PGCFL and PPGNet, our model is developed in Pytorch and operates on an Nvidia A100 GPU. Images are resized to 224x224, with data augmentation including random cropping, patch erasing, and flipping, adjusting orientation labels accordingly. Training ceases after 80 epochs for ATRW and YakReID-103, and 100 for ELPephants, starting with a fixed backbone for the initial three epochs. The learning rate is set at 0.001 for the backbone and 0.01 for other layers, reduced by tenfold after two-thirds of the epochs. SGD optimizer is used with momentum 0.9, weight decay $5e-4$, label smoothing 0.1, and a batch size of 30. Parameters d_{DVE}, λ_{reID}, and λ_{DVE} are set to 64, 2, and 0.2 respectively, employing Circle Loss with γ of 64 and m of 0.25.

For testing, features from the original and flipped images are concatenated for the Re-ID score, incorporating a re-ranking strategy for ATRW results. Despite not being discussed in the original works, we experimented with a vanilla triplet sampling strategy for batch processing, ensuring adequate negative samples for Circle Loss.

| | | | ATRW | | | YakReID-103 | | ELPephants | |
Methods	Org. Re-ID Task	Pose GT	mmAP	R@1(s)	R@1(c)	mAP	R@1	mAP	R@1
CLIP-Re-ID ViT [16]	Person	✗	56.5	86.3	72.0	49.8	82.2	12.7	25.3
PPGNet R-50 [18]	Animal	✓	68.3	81.2	81.1	-	-	-	-
ResNet50 [5]	Animal	✗	65.9	91.1	83.4	60.9	86.0	20.0	33.9
ViT [3]	Animal	✗	65.5	90.3	79.4	**61.3**	88.4	20.6	36.8
PGCFL [19]	Animal	✗	66.9	90.8	**86.3**	55.8	82.7	18.5	33.4
Ours	Animal	✗	**68.6**	**92.0**	84.6	61.0	**89.4**	**24.3**	**38.7**

Table 2. **Comparison to State-of-the-Art:** Our method is evaluated on three datasets, each representing a different species: ATRW (Tiger), YakReID-103 (Yak), and ELPephants (Elephants), originally proposed for various Re-ID tasks. Results from images with masked backgrounds, detailed in Sec. 2.1, are highlighted. Our model achieves top performance, surpassing existing baselines in mAP on ATRW and ELPephants, even outperforming PPGNet, which utilizes extra pose labels. For results on original images see supplementary material.

Effect of Background Removal As can be seen in Tab. 1 background removal significantly impacts performance, with a mmAP decrease of 10.2% when training and testing without backgrounds, versus a 19.2% drop when testing a full-image trained model on a masked background. However, models trained on masked backgrounds can still leverage backgrounds in testing, with an observed 9.7% mmAP increase, suggesting the importance of background-independent Re-ID models. Subsequent results are reported using masked backgrounds to emphasize identity recognition regardless of the background.

Comparison to State-of-the-Art Our method surpasses existing models like PGCFL [19], PPGNet [18] (ATRW challenge winner), CLIP-Re-ID [16] (a recent person Re-ID method), and baselines with ResNet50 and ViT backbones using $L_{base} = L_{ID} + L_{LR} + \lambda_{reID} \times L_{reID}$. Specifically, PPGNet, which depends on ground-truth poses, is evaluated solely on ATRW. Results can be found in Tab. A.4. Our approach leads in mmAP and R@1(s) on ATRW, equals ViT in mAP, and excels in R@1 on YakReID-103, while significantly outperforming all on ELPephants by a 5.8 mAP margin against PGCFL. CLIP-Re-ID's adaptation to animal datasets is hindered by greater intra-identity variations in animals, a challenge not as prevalent in human subjects. Although PCN-RERP shows promising mAP on YakReID-103, it's limited to datasets featuring animals in standing postures. Results without background masking are provided in supplementary material.

Qualitative Analysis of DVE. DVE helps the model embed body part information in its activation. Features for a given body part will share similarity across different entities. This helps align body parts and thus facilitates Re-ID. As shown in the two top rows of Fig. 4, when visualizing the descriptors $\Phi(\mathbf{x})$ learn through the loss L_{DVE} we can successfully match body parts.

Towards transfer between species. We also investigate the possibility of transfer between species. Using a model trained on tigers, we visualize third-layer features of the Re-ID backbone on yaks and elephants. While tigers, yaks, and elephants have widely different shapes and being only trained on tigers, our model was able to match body parts across species as seen in rows 3 and 4 of Fig. 4.

3.1. Limitations

Background removal. There are cases where both ISNet and SAM fail. In these cases, the combination of the two methods does not help. But these are rare cases and are usually linked to low image quality.

Inter-species transfer. While the proposed approach showed promising results on inter-species transfer, it also struggled in some specific cases. A model trained only on tigers has difficulty matching horns between yaks or tusks between elephants. We further illustrate these failure cases in the supplementary material.

4. Conclusion

We have introduced a novel method to advance animal re-identification. Our approach diverges from prior work by learning part-aware features in an unsupervised manner. Furthermore, by automatically masking the background, we not only address a recurrent challenge in animal Re-ID benchmarks but also effectively reduce the model's inclination to overly focus on the background. Lastly, we demonstrated that our approach represents an initial step towards the inter-species transferability of Re-ID models. In future studies, we plan to further refine our method's inter-species generalization capabilities by leveraging additional multi-species unsupervised training.

References

[1] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 330–346. Springer, 2020. 1

[2] Gong Cheng, Guangxing Wang, and Junwei Han. Isnet: Towards improving separability for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 1, 2

[4] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 852–860, 2022. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 2

[6] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7073–7082, 2018. 1

[7] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 1

[8] Lex Hiby, Phil Lovell, Narendra Patil, N Samba Kumar, Arjun M Gopalaswamy, and K Ullas Karanth. A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins. *Biology letters*, 5(3):383–386, 2009. 3, 1

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 2

[11] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 369–386. Springer, 2020. 1

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[13] Matthias Korschens and Joachim Denzler. Elpephants: A fine-grained dataset for elephant re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 4

[14] Lei Li, Tingting Zhang, Da Cuo, Qijun Zhao, Liyuan Zhou, and Suonan Jiancuo. Automatic identification of individual yaks in in-the-wild images using part-based convolutional networks with self-supervised learning. *Expert Systems with Applications*, 216:119431, 2023. 1, 2

[15] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 1, 3, 4

[16] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 1, 5, 2

[17] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 2

[18] Cen Liu, Rong Zhang, and Lijun Guo. Part-pose guided amur tiger re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 3, 5, 2

[19] Ning Liu, Qijun Zhao, Nan Zhang, Xinhua Cheng, and Jianing Zhu. Pose-guided complementary features learning for amur tiger re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 1, 2, 3, 5

[20] Shuang Liu, Wenmin Huang, and Zhong Zhang. Person re-identification using hybrid task convolutional neural network in camera sensor networks. *Ad Hoc Networks*, 97:102018, 2020. 1

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019. 2

[22] Julien Martin, Wiley M. Kitchens, and James E Hines. Importance of well-designed monitoring programs for the conservation of endangered species: Case study of the snail kite. *Conservation Biology*, 21, 2007. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 2

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 1

[26] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss:

A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020. 3

[27] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6361–6371, 2019. 1, 3

[28] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 1

[29] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10568–10577, 2020.

[30] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019.

[31] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1

[32] Tingting Zhang, Qijun Zhao, Cuo Da, Liyuan Zhou, Lei Li, and Suonan Jiancuo. Yakreid-103: A benchmark for yak re-identification. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1, 2, 3, 4

[33] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 1

Supplementary Material

In the supplementary material accompanying this paper, we offer enhanced visualizations that highlight the complexities involved in animal re-identification. Additionally, we conduct a comprehensive ablation study to assess the impact of each component within our proposed model. We also present qualitative results to demonstrate the model's performance, provide an in-depth analysis of the model's capability to transfer knowledge across different species, and furnish supplementary details regarding the implementation. Our aim is to provide a thorough understanding of our approach and its underlying mechanisms.

A.5. Challenge of Animal Re-ID

Telling apart different entities of the same animal species is a subtle task, as shown in Fig. A.5, one really needs to look at specific details of the distinguishing body part. For tigers, identification is usually based on body stripe patterns [8]. For yaks, it is mostly their horns, sometimes fur and texture can help. For elephants, ears and tusks are important for identification. Matching the same body part between different entities is key to improving the Re-ID performance.

Figure A.5. **Sample images from YakReID-103 and ATRW and ELPephants**. In each row, the first two images are of the same entity while the last image is of another.

A.6. Ablation study.

To understand which element of our model contributes to its overall performance we conduct an extensive ablation study, Tab. A.3 on ATRW, to test the impact of each individual loss (L_{DVE}, L_{ID}, L_{reID}, L_{LR}) as well as the

Model Components					ATRW	
L_{DVE}	L_{ID}	L_{ReID}	L_{LR}	B.S.	mAP	R@1(c)
	✓				54.6	74.9
		✓			56.3	72.0
	✓	✓			54.4	69.7
	✓		✓		61.4	81.1
		✓	✓		60.5	76.6
	✓	✓	✓		63.4	78.9
✓	✓	✓	✓		63.4	78.3
	✓	✓	✓	✓	63.2	80.6
✓	✓	✓	✓	✓	64.9	82.9

Table A.3. **Ablation study.** We study the influence of 5 different components of our approach: The impact of our 4 objective functions, and the batch sampling strategy (B.S.).

batch sampling strategy. The metric-based loss L_{reID} performs better than the classification loss L_{ID} as the former learns more generalized feature representation. The use of orientation loss L_{LR} in conjunction with L_{reID} and L_{ID} consistently boosts the performance, suggesting the importance of the pose information. L_{DVE} loss and batch sampling, together, further improve the mAP score by 2.4%. This makes it evident that DVE loss helps to leverage animal part-specific information for better Re-ID. All results reported in this ablation study are without re-ranking.

A.7. Detailed Comparison to State-of-the-Art

Baselines. We compare our approach to PGCFL [19], PPGNet [18] the winner of the ATRW challenge, CLIP-Re-ID [16] a recent person-reid method, and two other baselines: First a simple baseline using a ResNet50 backbone trained with $L_{base} = L_{ID} + L_{LR} + \lambda_{reID} \times L_{reID}$. Then, a baseline using a more modern ViT [3] backbone is trained with L_{base} as well. PPGNet is run for 160 epochs and we followed the experimental settings from the original work. For the two training stages of CLIP-Re-ID, the ResNet-based model run for 120 and 60 epochs, and the ViT-based model run for 60 and 80 epochs. The rest of the experimental settings adhere to the CLIP-Re-ID original work. For all other methods, we use the same experimental setting previously described. PPGNet relies on groundtruth poses

Methods	Org. Re-ID Task	Pose GT	ATRW			YakReID-103		ELPephants	
			mmAP	R@1(s)	R@1(c)	mAP	R@1	mAP	R@1
			Masked Backgrounds						
CLIP-Re-ID R-50 [16]	Person	✗	54.4	82.6	71.4	48.2	85.6	13.1	25.5
CLIP-Re-ID ViT [16]	Person	✗	56.5	86.3	72.0	49.8	82.2	12.7	25.3
PPGNet R-50 [18]	Animal	✓	68.3	81.2	81.1	-	-	-	-
ResNet50 [5]	Animal	✗	65.9	91.1	83.4	60.9	86.0	20.0	33.9
ViT [3]	Animal	✗	65.5	90.3	79.4	**61.3**	88.4	20.6	36.8
PGCFL [19]	Animal	✗	66.9	90.8	**86.3**	55.8	82.7	18.5	33.4
Ours	Animal	✗	**68.6**	**92.0**	84.6	61.0	**89.4**	**24.3**	**38.7**
			Original Backgrounds						
CLIP-Re-ID R-50 [16]	Person	✗	63.6	94.0	81.7	41.4	82.2	3.7	10.8
CLIP-Re-ID ViT [16]	Person	✗	61.4	90.0	83.4	50.0	83.7	3.9	12.6
PPGNet R-50 [18]	Animal	✓	77.9	99.4	90.8	-	-	-	-
PGCFL [19]	Animal	✗	74.5	**95.7**	**90.3**	64.3	91.8	10.7	24.7
PCN-RERP [14]	Animal	✗	-	-	-	**68.6**	91.8	-	-
Ours	Animal	✗	**76.2**	**95.7**	88.0	66.1	**92.3**	**14.1**	**30.3**

Table A.4. **Comparison to State-of-the-Art:** Our method is evaluated on three datasets, each representing a different species: ATRW (Tiger), YakReID-103 (Yak), and ELPephants (Elephants), originally proposed for various Re-ID tasks. Results from images with masked backgrounds are highlighted. Our model achieves top performance, surpassing existing baselines in mAP on ATRW and ELPephants, even outperforming PPGNet, which utilizes extra pose labels. When considering original images, our model outperforms PGCFL across all datasets and matches PPGNet on ATRW. CLIP-Re-ID, designed for person Re-ID, fails to generalize to animals due to high intra-class variations. PCN-RERP performs well on Yak dataset but lacks generalization to non-standing animal postures.

and therefore the evaluation is only shown on the ATRW dataset. Additionally, we compared with PCN-RERP [14] who report their method on YakReID-103 dataset.

Results. In Tab. A.4, our approach outperforms prior state-of-the-art models on the ATRW dataset for mmAP and R@1(s). On the YakReID-103, it matches the ViT baseline in mAP, and outperforms all in terms of R@1. On the ELPephants dataset, it significantly outperforms all baselines with a margin of 5.8 mAP points w.r.t. PGCFL. CLIP-Re-ID baseline originally proposed for person re-id task fails to generalize on the animal dataset. This can be attributed to the fact of higher intra-identity variations occurring in animals than in persons, which CLIP-based models fail to capture. CLIP-based models are better known for their zero-shot inter-class/identity classification. For completeness, we also provide results on the original benchmark without masking backgrounds. Here, overall metrics are higher but the ranking of the different baselines is similar. PCN-RERP [14] has better overall mAP on Yak dataset but this approach cannot be generalized to dataset where animals are in the non-standing posture.

A.8. Background removal criterion

In the proposed protocol to remove the background, we aggregate the object masks generated by SAM if they satisfy a criterion $h(\mathbf{m}; S_{ISNet}(\mathbf{x})) \geq \Theta$. The function h is manually tuned to better fit each dataset. We list the criteria used for each dataset here.

For ATRW we used:

$$h(\mathbf{m}; S_{ISNet}(\mathbf{x})) = \frac{|\mathbf{m} \cap S_{ISNet}(\mathbf{x})|}{|\mathbf{m}| \cup |S_{ISNet}(\mathbf{x})|} \geq 0.3 \quad (9)$$

For ELPephants we used:

$$h(\mathbf{m}; S_{ISNet}(\mathbf{x})) = \frac{|\mathbf{m} \cap S_{ISNet}(\mathbf{x})|}{|\mathbf{m}|} \geq 0.5 \quad (10)$$

For YakReID-103 we directly used the output from SAM predictor, no filtering was applied.

A.9. Additional Qualitative Results

We provide further results for transfer between species. In Fig. A.6 we show that the local descriptors learned by our approach also work for models trained on elephants or yaks,
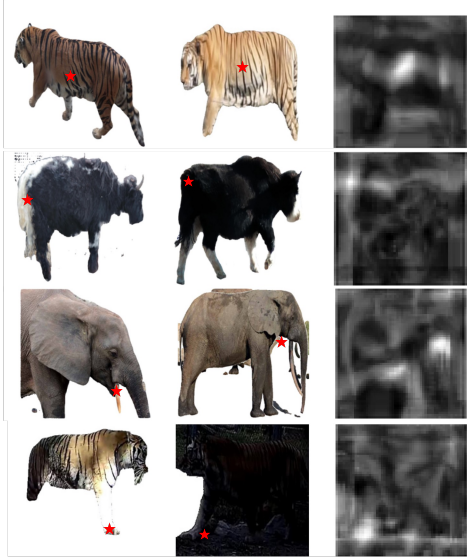
Figure A.6. **Intra-Species visualization of SEResNet trained with DVE loss.** In each row, the red point from the left-most image is queried in the second image, and the red point in the middle image is its matching point, the right-most image is a heatmap of cosine similarities between the middle image and the red point in the left-most image. Interestingly, the last row show that even when images are of bad quality and of different views, the matching can be good.
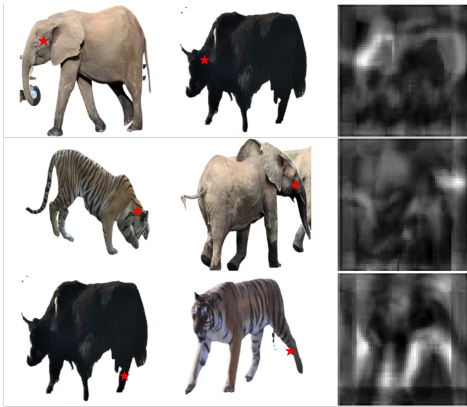

Figure A.7. **Inter-Species visualization of SEResNet trained with DVE loss.** In each row, the red point from the left-most image is queried in the second image, and the red point in the middle image is its matching point, the right-most image is a heatmap of cosine similarities between the middle image and the red point in the left-most image.

even in challenging scenarios with low-quality images (and failed background removal). In Fig. A.7 we provide additional results for the matching of local DVE features across species. Finally in Fig. A.8 we compare this transfer capability with two baselines: PGCFL and ResNet50. In both cases, they fail to match body parts, confirming that the
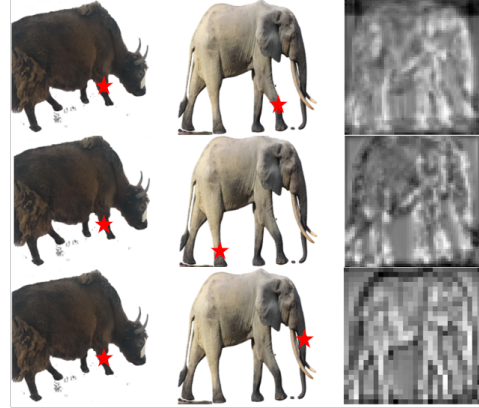

Figure A.8. **Visualization of yak-elephant body part matching using features from the third layer of different models trained on tigers.** In each row, the red point from the left-most image is queried in the second image, and the red point in the middle image is its matching point, the right-most image is a heatmap of cosine similarities between the middle image and the red point in the left-most image. From top to bottom, each row correspond to our model, PGCFL and Resnet50 respectively.

matching capacity of our model is indeed coming from the use of L_{DVE}.

A.10. Transfer Quantitative Evaluation

We provide a quantitative evaluation of transfer between species. Results can be found in Tab. A.5. While performance on transfer is lower than the supervised model, the model still manages to transfer some knowledge across species. Note that those performances are obtained on the masked background so the model can only rely on the animal's appearance.

A.11. Hyperparameters

Our method relies on two main hyperparameters, λ_{DVE} and λ_{reID}. For λ_{DVE} we tested values in the range $[0, 2]$ with 0.2 increments and settled on the value 0.2. For λ_{reID} we tested values $1, 2, 5$ and settled for 2.

A.12. Code

The code is provided in the following link: `https://github.com/Chloe-Yu/Animal-Re-ID`. It contains all the scripts necessary to generate the masked background benchmarks and train and evaluate the proposed model.

A.13. Data

We provide the complete heading direction annotation for the ELPephant dataset in

Training Data	ATRW			YakReID-103		ElPephants	
	mmAP	R@1(s)	R@1(m)	mAP	R@1	mAP	R@1
ATRW	68.6	92.0	84.6	35.9	73.1	6.7	17.6
YakReID-103	49.5	80.9	70.9	61.0	89.4	6.5	16.3
ElPephants	47.2	80.9	66.3	34.5	75.0	24.3	38.7

Table A.5. **Evaluation of inter-species transferability** We propose to evaluate the transferability of the proposed approach between three species: Tiger, Yak and Elephant. Each row of the table correspond to our model trained on a single species (Training data) and evaluated on the test set of the three species. While the performance on transfer are below than the fully supervised one, the proposed model is able to transfer meaningful representation between species.

re_mapped_filtered_anno_elephant.csv together with our code.